Metode Klasifikasi Untuk Deteksi Uniform Resource Locator (URL) Berdasarkan Jenis Serangan Menggunakan Algoritma Naive Bayes, C4.5 dan K-Nearest Neighbor

Moh Yunus, Dwi Widiastuti, Hasma Rasjid dan Yulia Chalri
Sistem Informasi, Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Gunadarma
Jl. Margonda Raya No. 100, Depok, Jawa Barat 16424
{sunuy165, dwidiastuti, hasmapsa, liapsa}@staff.gunadarma.ac.id

ABSTRAK

Keamanan sistem merupakan hal yang penting dalam menjaga integritas dan kerahasiaan data. Kebocoran data dapat mengancam seiring dengan meningkatnya sumber daya manusia.Banyak pengembang sistem yang berhasil melindungi sistemnya dariURL jahat yang telah dikenal, tentu ini hanya menyelesaikan sebagian dari masalah yang ada, karena URL jahat yang tidak dikenal atau URL jahat baru menjadi ancaman baru dalam masalah ini. Situs web peringkat yang terpercaya yaitu Alexa telah menyampaikan bahwa banyak URL palsu yang dapat dikompromikan, yang mana ini merupakan hal sulit bagi pengembang untuk dapat membedakan atau mengklasifikasikan URL jahat berdasarkan jenis serangannya. Penggunaan algoritma klasifikasi dalam hal ini K-Nearest Neighbor (K-NN), C4.5 dan Naive Bayes merupakan pilihan penting dalam menentukan penggelompokan berdasarkan jenis serangan dengan bantuan analisis leksikal yang baik dan efektif dalam mendeteksi sistem proaktif pada URL yang terbagi dalam 4 jenis yaitu phising, spam, malware, dan defacement. Pendekatan leksikal pada proses klasifikasi dengan algoritma K-NN, C4.5 dan Naïve Bayes mampu mengelompokkan URLs Jahat berdasarkan jenis serangannya tingkat akurasi keberhasilan rata-rata diatas 90% dari dataset yang telah digunakan.

Kata Kunci : *Klasifikasi, K-NN, C4.5, Naïve Bayes, Leksikal Analisis*

PENDAHULUAN

Pemahaman dan kesadaran yang kurang terhadap isu keamanan sistem dapat mengancam setiap saat khususnya bagi para pengembang. Kebocoran data atau perusakan dapat mengancam seiring dengan meningkatnya sumber daya manusia. Dewasa ini jutaan situs web jahat semakin banyak dengan berbagai model penipuan termasuk pemasaran barang palsu, melakukan penipuan keuangan (misalnya, "*Phishing*") dan menyebarkan *malware* (misalnya, melalui eksploitasi). Semua kegiatan ini diawali dari pengaksesan *Uniform Resource Locator* (URL) atau link yang sudah dipasang file berbahaya didalamnya, karena ini merupakan vector atau alamat yang membawa pengguna internet ke dalam jaringan tertentu.

Google sebagai situs populer telah mendeteksi ribuan situs berbahaya setiap harinyaberdasarkan laporan dari *Safe-Browsing*yang merupakan indikasi yang jelas dan bukti bahwa web phising atau website jahat digunakan oleh para penjahat untuk melakukan aksinya. web phising digunakan untuk mencuri informasi pribadi, seperti kartu kredit dan *password* dan mengelabui pengguna dengan berbagai cara untuk memikat hati pengguna agar mengunjungi situs atau URL perangkap.

Kunjungan ke situs web yang terinfeksi *malware* atau sejenisnya memungkinkan penyerang dapat mendeteksi kelemahan atau melakukan kejahatan bahkan ancaman di setiap pengguna aplikasi dan memaksanya untuk mengunduh banyak file binari perangkat lunak perusak agar dapat dikontrol oleh penyerang. Serangan *malware* saat ini memungkinkan penyerang dalam mengontrol penuh sistem yang dikendalikan dari jarak jauh baik itu informasi yang bersifat sensitive atau instalasi utilitas.

Serangan jahat dengan media URL cenderung memiliki masa hidup yang lebih pendek dibanding dengan URL baik atau normal, mereka sering muncul dengan konten yang berbeda kadang mereka berisi iklan atau info lain yang menarik agar target masuk dalam perangkapnya. URL sebagai titik masuk sebuah sistem, berfungis sebagai kontrol pengguna dengan tujuan dapat dikendalikan oleh server *malware*yang telah

dibuat. *Malware* seringkali mengubah konten URL untuk mengelabuhi pengguna sehingga sangat sulit untuk dideteksi. oleh karena itu, para pengembang berupaya untuk mengatasi masalah kemanan ini yang menfokuskan pada daftar url yang jahat.

Banyak pengembangyang berhasil melindungi sistemnya dan pengamankan pengguna dari pengaksesan URL jahat yang telah dikenal, tentu ini hanya menyelesaikan sebagian dari masalah yang ada, karena URL jahat yang tidak dikenal atau URL jahat baru yang muncul diseluruh seluruh sistem umumnya menjadi yang terdepan atau ancaman baru dalam masalah ini. Situs web peringkat yang terpercaya yaitu *Alexa* telah menyampaikan banyak URL palsu yang dapat dikompromikan biasanya disebut defacement URL. Mengeksplorasidan mengkategorikan URL berbahaya sangat penting guna membantu dalam mengkategorikan jenis serangan dari URL jahat itu sendiri. Oleh sebab itu, dibutuhkan suatu analisis leksikal yang baik dan efektif untuk deteksi sistem proaktif pada URL. Diketahui bahwa yang terbaru dari sebuah URL adalah teknik *obfuscation*yang mana teknik ini menjadi teknik yang sangat efektif untuk strategi jebakan.

Berdasarkan permasalahan ini, sangat diperlukan metode penyelesaian yang efektif dalam proses pengklasifikasian agar dapat memilih dan memilah berdasarkan jenis dan kategori yang ada, agar dapat mempermudah dalam menanggulangi permasalahan. Solusi dari permasalahan ini adalah dibutuhkan sebuah*machine learning*yang mampu memahami dan melakukan proses pengelompokkan dari hasil data yang didapat agar nantinya dapat menjadi alat bantu dalam mendeteksi dan mengklasifikasikan jenisURLs berdasarkan serangannya.

Mengidentifikasi URLsperlu sebuah sebuah analisis leksikal untuk dapat memahami secara mendalam URLs yang terdapat aktivitas jahatagar nantinya dapat dikategorikan kedalam kelompok sesuai dengan maksud dan tujuan dari serangan itu sendiri. Jenis URL yang telah dikumpulkan dari berbagai situs web jenis URL tersebut adalah *Benign URLs, Spam URLs, Phising URLs, Malware URLs dan Defacement URLs.*Teknik*obfuscation* digunakan sebagai metode yang umum dalam menyamarkan URL jahat yang mana URL yang dihasilkan dari teknik ini terkadang tidak masuk dalam kelima kategori tersebut, sehingga dapat menyulitkan dalam proses pengklasifikasiannya.

Penggunaan analisis statis pada fitur leksikal dengan maksud menjadikan URL ini termasuk dalam kategori yang baik atau jahat belum mendapatkan hasil maksimal. oleh karen itu, dibutuhkan sebuah algoritma klasifikasi untuk mendapatkan hasil maksimal dalam mengelompokkan URLs berdasarkan jenis serangannya. Proses pengelompokkan ini juga dibantu dengan pendekatan leksikal analisis agar mendapatkan hasil yang maksimal.

Proses pengelompokkan ini sangatlah penting bagi pengembang khususnya untuk dapat memahami bahwa tidak setiap URLs yang tidak dikenali merupakan URLs Jahat. Oleh karena itu diperlukan beberapa algoritma *machine learning*dalam proses pengelompokan ini, dalam hal ini penulis menggunakan tiga algoritma untuk melakukan proses pengklasifikasian agar dapat dijadikan pembanding hasil diantara ketiga metode tersebut yaitu Naive Bayes, C4.5 dan K-Nearest Neighbor (KNN).Metodeklasifikasi dengan berbagai macam algoritma telah banyak digunakan pada proses pengklasifikasian akan tetapi hasil yang didapatkan tidak selalu maksimal. Penambahan pendekatan leksikal analisis merupakan kombinasi yang baik untuk mendapatkan hasil maksimal dalam proses pengklasifikasian. Ini merupakan solusi jitu dalam membantu proses pengelompokan.

Tujuan dari Penelitian ini adalah Mampu mengelompokkan URLs Jahat berdasarkan analisa leksikal pada proses klasifikasi dengan pendekatan algoritma*machine learning*dalam hal ini adalah Naive Bayes, C4.5 dan K-Nearest Neighbor (KNN).Jumlah Data yang akan digunakandiambil dari berbagai sumber yaitu *Alexa, WEBSPAM-UK2007, DNS-BH,*dan *Repository Of Active Phishing Sites* dengan jumlah data yang diambil sekitar 114.400 URLs yang diambil dari *Canadian Institute For Cybersecurity.***[13]**Berdasarkan

dengan jumlah yang besar ini, dibutuhkan sebuah untuk proses pengklasifikasian yang efisien dengan bantuan analisa leksikal yang efektif dalam proses memahami dan mengenali jenis serangan agar dapat diketahui dari data tersebut mana yang termasuk dari kelima kategori tersebut.

TINJAUAN PUSTAKA

Analisis Leksikal

Menurut Mamun**,**fitur leksikal adalah sifat tekstual suatu URL dimana Panjang dari hostname dan panjang URL dapat dideskripsikan berdasarkan pola URL sehingga ini dapat menghasilkan karakter input dalam kode untuk mendapatkan sebuah token**.[8]**Sedangkan menurut Michael Darling dan Greg**,** adalah model Bahasa standar yang bertujuan menghitung probabilitas kesamaan nilai dan sifat-sifat dari URLs. Jika URL yang diberikan mengandung token yang ditemukan berdasarkan elemen terkait maka diberi nilai 1 begitu juga sebaliknya**.[9]**

K-Nearest Neighbor (K-NN)

Tujuan dari algoritma ini adalah mengklasifikasikan obyek berdasarkan *attribut* dan *training sample*. Clasifier tidak menggunakan bantuan apapun untuk dapat mencocokkan dan hanya berdasarkan pada memori**.[11]**Algoritma K-NN merupakan algoritma yang menentukan nilai jarak pada pengujian data testing dengan data training berdasarkan nilai terkecil dari nilai ketetanggaan terdekat. Jarak yang digunakan adalah jarak *Euclidean Distance*. Jarak Euclidean adalah jarak yang paling umum digunakan pada data *numeric*.**[5]**

Naive Bayes

Naive Bayes Classifier atau sering disebut *Bayesian Classification* adalah metode pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class.**[2]**

Pengklasifikasian terdapat 2 proses yang dilakukan yaitu :

1. Proses *training*
   Pada proses ini dilakukan training set yang sudah diketahui label-labelnya untuk membangun model.
2. Proses *testing*
   Proses ini intuk mengetahui keakuratan model yang dibangun pada proses training, umumnya digunakan data yang disebut test set untuk memprediksi label.**[2]**

C4.5

Algoritma *decision tree*digunakan untuk membangun sebuah pohon keputusan yang mudah dimengerti,fleksibel, dan menarik karena dapat divisualisasikan dalam bentuk gambar**.[1]** Metode ini berfungsi untuk mengubah fakta menjadi pohon keputusan yang merepresentasikan aturan yang dapat mudah dimengerti dengan bahasa alami. Algoritma ini sudah sangat terkenal dan disukai karena memiliki banyak kelebihan.

Kelebihan ini misalnya dapat mengolah data numerik dan diskret, dapat menangani nilai atribut yang hilang, menghasilkan aturan-aturan yang mudah diinterpretasikan dan performanya merupakan salah satu yang tercepat dibandingkan dengan algoritma lain.**[4]**

Obfuscations URLs

Obfuscated adalah teknik menyamarkan kode dengan tetap memelihara semantik (isi) data, dengan tujuan agar tidakdengan mudah dibaca orang lain.**[3]***Obfuscation*bertujuan untuk menyesatkan atau mengelabui penyelidik dengan menyembunyikan atau menghapus bukti tentang sumber dan sifat serangan.**[12]**.

Waikato Environment for Knowledge Analysis (WEKA)

WEKA adalah sebuah paket *tools machine learning*praktis. WEKA merupakan singkatan dari *WaikatoEnvironment for Knowledge Analysis*, yang dibuatdi Universitas Waikato, New Zealand untukpenelitian, pendidikan dan berbagai aplikasi. WEKAmampu menyelesaikan masalah-masalah datamining di dunia nyata, khususnya klasifikasi yangmendasari pendekatan-pendekatan *machine learning*.Perangkat lunak ini ditulis dalam hirarki *class Java*dengan metode

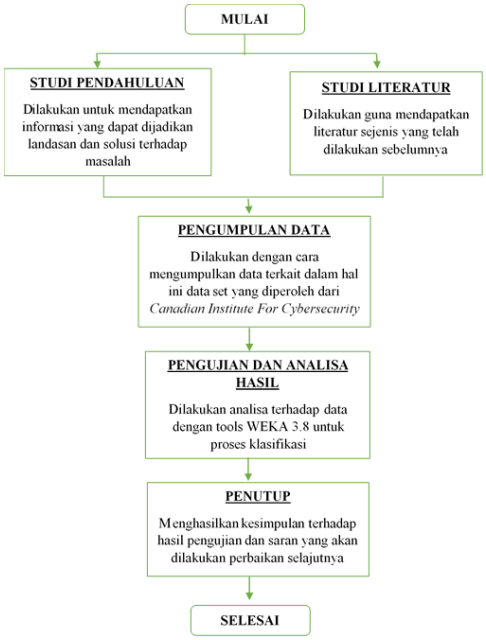
berorientasi objek dan dapat berjalanhampir di semua *platform*.**[6]**

Tahapan *preprocessing* datadilakukan menggunakan tools WEKA 3.8.1.Tahapan-tahapan prepocesing yang dilakukansebagai berikut:

1. *Lower Case Tokens* berfungsi untukmembuat data tweet menjadi huruf kecil semua.
2. *Normalization* dilakukan untukmenormalkan kata-kata yang tidak baku.
3. *Tokenization* dilakukan untuk memecahmenjadi beberapa kata atau kumpulan kata yang berdiri sendiri.
4. *Cleansing* yaitu proses menghapus symbol-simbol yang kurang penting dalam data.
5. *Filtering* dilakukan untuk menghapus katakata yang kurang penting atau kurangberpengaruh terhadap proses klasifikasi nantinya.**[10]**

METODE PENELITIAN

Obyek yang dijadikan pengujian pada penelitian ini merupakan sekumpulan data sekunder yang diambil dari *University Of New Brunswick Canadian Institute For Cybersecurity* dimana institusi ini telah menyediakan data set yang dapat digunakan oleh Universitas, swasta atau peneliti independen untuk dijadikan sebagai bahan pengujian. Data set yang dipakai pada penelitian ini merupakan sekumpulan data set URLsyang sudah disediakan oleh institusi untuk umum dengan total jumlah data yang terkumpula adalah sekitar 114.250 data set tanpa adanya proses pemilihan. Adapun skema metode penelitian dapat digambarkan pada Gambar 1.



Gambar 1. *Skema Penelitian*

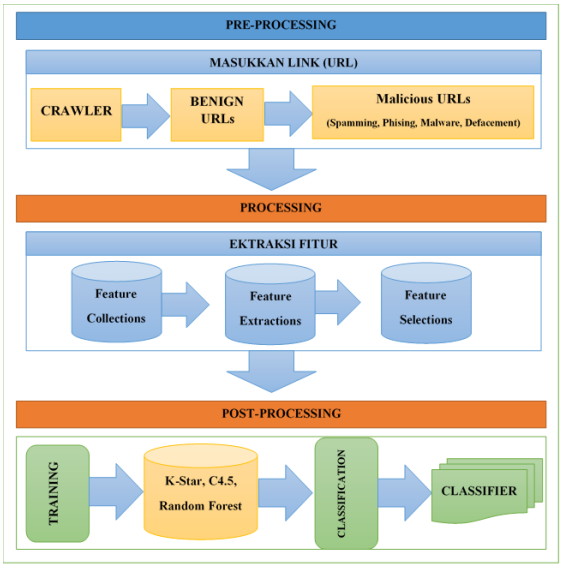
Tahapan pertama pada penelitian ini dimulai dengan studi pendahuluan yaitu peneliti menguraikan permasalahan dan memberikan solusi terhadapa permasalahan yang terjadi dalam sebuah sistem informasi serta melakukan studi pustaka mengenai referensi yang berkaitan dengan permasalahan yang ada.

Tahap kedua adalah teknik pengumpulan data yang terkait dengan penelitian dalam hal ini data set yang telah dihimpun oleh *Canadian Institute For Cybersecurity d*ari berbagai sumber untuk selanjunya dijadikan sebagai bahan penelitian. Pada tahapan ini peneliti tidak melakukan proses pengkajian ulang atau proses triger atau evaluasi dari data set yang telah disediakan untuk menjaga kemurnian data.

Tahapan selanjutnya adalah proses analisa terhadap data yang didapatkan dengan menggunakan *tools pembantu* dan analisa berdasarkan pendekatan yang telah disebutkan sebelumnya, dalam hal ini tools yang dipakai adalah WEKA 3.8 untuk proses analisa dan perbandingan data.

Secara Garis Besar Penelitian ini menggunakan metode observasi dan literasi untuk proses pendekatan ilmiah, dengan analisa kuantitatif pada proses pengujian, terdapat tiga proses utama dalam penelitian

ini yaitu *Pre-Processing*, *Processing* dan *Post-Processing* untuk mendapatkan hasil yang maksimal seperti pada Gambar 2



Gambar 2. *URLs Klasifikasi Arsitektur*

Gambar 2 menjelaskan bahwa untuk tahap *Pre-Processing* adalah tahap pencarian data dengan proses *crowling* setiap URLs tanpa harus memilah, dan selanjutnya akan diproses dengan analisa leksikal untuk memilah URLs Jahat dan Tidak Jahat. Pada Proses Selanjutnya URLs Jahat dikelompokkan sendiri untuk proses klasifikasi berdasarkan jenisnya. Pada Tahap *Processing* adalah Proses Ektraksi Fitur yaitu menyeleksi dan memilah dengan proses leksikal untuk memahami URLs berdasarkan jenis serangan. Pada tahap Terakhir adalah *Post-Processing* dimana tahap ini adalah Tahap Pengklasifikasian dengan *machine learning* dalam hal ini K-NN, C4.5, dan *Naive Bayes.* dengan data training yang telah dianalisa sebelumnya dengan analisa leksikal.

HASIL DAN PEMBAHASAN

Sekitar 114.000 dataset yang diambil dari berbagai sumber dibagi menjadi 2 bagian yaitu URLs jahat dan Tidak Jahat dengan Proses Klasifikasi Pada URLs Jahat kedalam 4 bagian yaitu

1. *Benign URLs* Sekitar 35.000 dari *Alexa*.com.
2. *Spam URLs* Sekitar 12.000 dari *Public Spam Dataset*[14].
3. *Phising URLs* Sekitar 10.000 dari *Repository Phising Sites*[10].
4. *Malware URLs* Sekitar 11.000 dari *DNS-BH*[7].
5. *Defacement URLs* 25.000 dari *Zone-H*[16].

PENGUJIAN DAN VALIDASI

Tahapan ini dilakukan evaluasi dan strategi pencarian untuk menemukan fitur yang signifikan dengan analisis yang mendalam. Pada proses validasi data penelitian ini menggunakan 10 *Fold Cross-Validation*untuk melihat tingkat presisi dan kebenaran.

Pengujian	Dataset									
1	■									
2		■								
3			■							
4				■						
5					■					
6						■				
7							■			
8								■		
9									■	
10										■

Gambar 3. *Ilustrasi 10 Fold Cross Validation*

Melakukan evaluasi performa *TP rate, FP rate, Precision, Recall* dan *F-measure* darieksperimen yang telah dilakukan. Evaluasi dilakukan dengan menggunakan *Confusion Matrix* yaitu *true positive rate (TP rate), true negative rate (TN rate), false positive rate (FPrate)* dan *false negative rate (FN rate)* sebagai indikator.

TP rate adalah persentase dari kelas positif yang berhasil diklasifikasi sebagai kelaspositif, sedangkan *TN rate* adalah persentase dari kelas negatif yang berhasil diklasifikasi sebagai kelas negatif.

FP rate adalah kelas negatif yang diklasifikasi sebagai kelas positif. *FN rate* adalah kelas positif yang diklasifikasi sebagai kelas negatif.

$$Precision\ (pr) = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}\ Recall\ (rc) = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$F = \frac{2\text{PR}}{\text{P} + \text{R}}$$

Pencapaian hasil Pengujian yang effisien digunakan parameter acak (*random tree*) untuk hasil yang lebih baik dan jauh dari kesalahan.

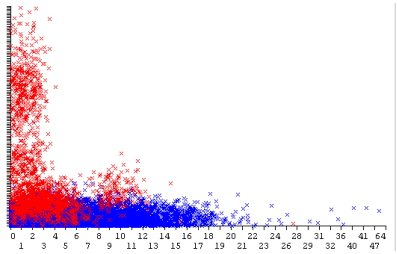
ANALISA DAN HASIL

Berdasarkan analisa sebelumnya bahwa dengan pendekatan leksikal untuk mengenali URL berbahaya dan juga menganalisa URL *obfuscation* terhadap jenis serangan maka hasil pengklasifikasian dengan pendekatan *machine learning* dalam hal ini adalah *K-NN, C4.5,* dan *Naive Bayes* dapat dilihat pada Tabel 1dengan fitur leksikal.

Tabel 1. *Fitur Leksikal*

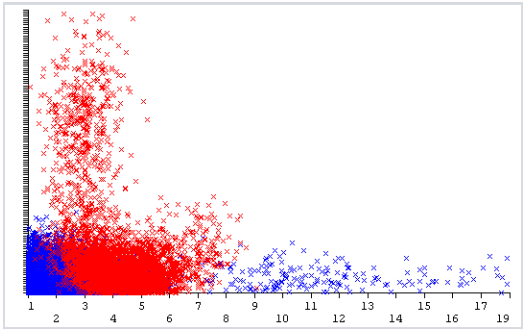
Dataset	Algoritma	Hasil	
		Pr	Rc
Phising	K-NN	0,98	0,95
	C4.5	0,94	0,91
	Naive Bayes	0,96	0,92
Spam	K-NN	0,98	0,99
	C4.5	0,98	0,98
	Naive Bayes	0,93	1
Malware	K-NN	0,98	0,98
	C4.5	0,96	0,97
	Naive Bayes	0,91	0,8
Defacement	K-NN	0,99	0,99
	C4.5	0,98	0,98
	Naive Bayes	0,98	0,93

Berdasarkan Tabel 1 dapat disimpulkan bahwa penggunaan algoritma K-NN, C4.5dan *Naive Bayes*untuk tingkat *precision* dan *recall* memperoleh hasil maksimal rata-rata 0,9 untuk masing-masing algoritma.



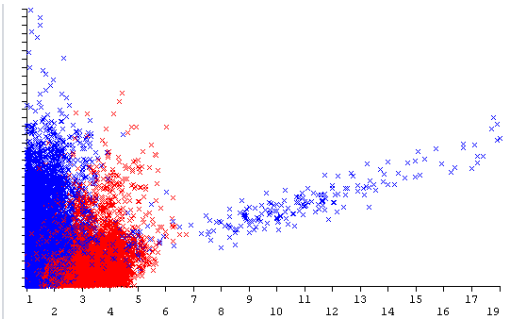
Gambar 4. *Klasifikasi Phising URLs*

Berdasarkan Gambar 4 dapat dijelaskan bahwa yang berwarna merah adalah *Phising URLs* dan yang berwarna biru adalah *Benign URLs* dimana proses klasifikasi dari URLs*Phising* dan *Benign URLs* telah berhasil dengan tingkat *precison* rata-rata 0,96untuk masing-masing algoritma sedangkan *recall* mencapai rata-rata 0,93 untuk masing-masing algoritma.



Gambar 5. *Spam URLs*

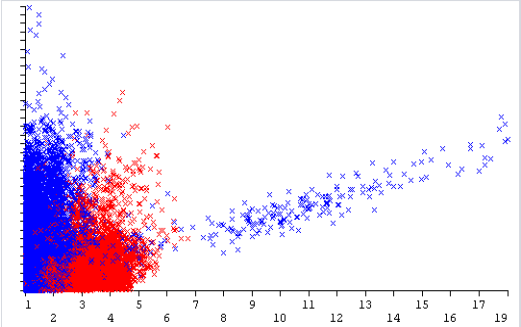
Berdasarkan visualisasi dari Gambar 5 dapat dijelaskan bahwa untuk warna merah adalah proses *Spam URLs* sedangkan untuk warna biru adalah proses *Benign URLs* dimana dapat disimpulkan bahwa untuk setiap algoritma pada proses pengklasifikasian berhasil mencapai tingkat *precision*rata-rata 0,96untuk setiap algoritma sedangkan *recall* mencapai rata-rata 0,99 untuk masing-masing algoritma.



Gambar 6. *Malware URLs*

Berdasarkan Gambar 6 dapat dijelaskan bahwa proses klasifikasi dari *URLs Malware*terhadap *Benign URLs*berhasil dengan tingkat *precision* rata-rata 0,95 sedangkan *recall* untuk masing-masing algoritma adalah0,91

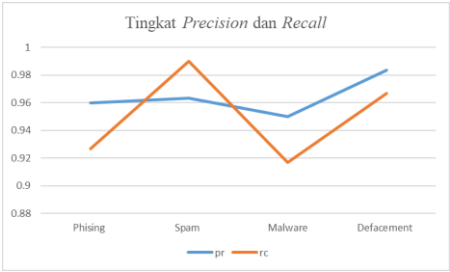
dimana warna merah adalah *Malware URLs* sedangkan warna biru adalah *Benign URLs*.



Gambar 7. *Defacement URLs*

Berdasarkan Gambar 7 dapat dijelaskan bahwa untuk warna merah adalah proses *Defacement URLs* sedangkan untuk warna biru adalah proses *Benign URLs*disimpulkan bahwa proses klasifikasi berhasil dengan tingkat *precision*rata-rata 0,98 untuk masing-masing algoritma sedangkan rata-rata recall adalah 0,96 untuk masing-masing algoritma.

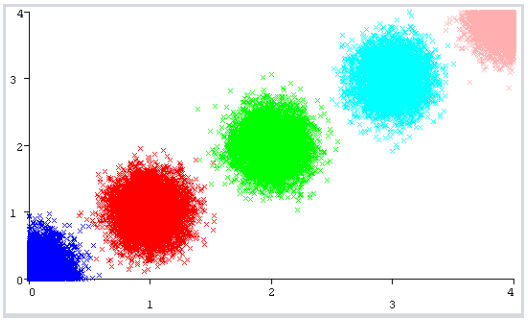
Tingkat *precision* dan *recall* untuk masing-masing algoritma dapat dilihat pada gambar 8 berikut ini.



Gambar 8. *Tingkat Precision dan Recall*

Berdasarkan gambar 8 dapat disimpulkan bahwa tingkat precision dan recall untuk masing-masing kategori menunjukkan tingkat kisaran 0,9 sampai 0,99 ini menunjukkan bahwa hampir disemua algoritma bisa membuktikan proses klasifikasi dengan baik meskipun untuk data *malware URLs* ada sedikit penurunan pada proses recall akan tetapi precision masih ada dikisaran 0,9.

Gambaran proses keselurahan semua data yang telah dilakukan proses klasifikasi berdasarkan jenis serangan dapat dilihat seperti pada Gambar 9.



Gambar 9. *Proses Klasifikasi Data Pada Semua Jenis Serangan*

Berdasarkan Gambar 9 dapat dijelaskan bahwa proses klasifikasi dari keseluruhan jenis serangan berhasil dengan maksimal, setiap jenis serangan dapat dipisahkan berdasarkan empat jenis serangan yaitu *Phising, Spam, Malware,* dan *Defacement.*Penjelasan warna pada Proses adalah dimana untuk warna biru adalah proses *Benign URLs* sedangkan untuk warna merah adalah *Phising URLs*, untuk warna hijau adalah *Spam URLs*, warna *tosca* adalah proses *Malware URLs* dan yang terakhir adalah warna *Pink* adalah proses *Defacemanet URLs.*

PENUTUP

Berdasarkan hasil pengujian algoritma K-NN, C4.5 dan *Naive Bayes*terhadap data yang ada maka dapat disimpulkan bahwa pengklasifikasian URLs jahat berdasarkan jenis serangannya telah berhasil dilakukan. Tingkat keberhasilan dari ketiga metode tersebut diatas 90% baik itu untuk tingkat presisi dan akurasi dengan*ten-fold cross-validation* pada setiap tahapan algoritmanya.

Meskipun pada hasil analisis *malware URLs* pada proses *recall*dibawah 90% akan tetapi sudah mampu mengklasifikannya kedalam jenis serangan yang sudah ada. Untuk penelitian selanjutnya sebaiknya dilakukan uji coba dataset terbaru untuk jenis serangan model terbaru.

DAFTAR PUSTAKA

[1] . Abdul Rohman, Vincent Suhartono, Catur, Supriyanto, "*Penerapan Algoritma C4.5 Berbasis Adaboost Untuk Prediksi Penyakit Jantung*"Jurnal Teknologi Informasi, Volume 13 Nomor 1, Januari 2017, ISSN 1907-3380, http://research. pps. dinus. ac. Id

[2] . Akhmad Pandhu Wijaya, Heru Agus Santoso,"*Naive Bayes Classification Pada Klasifikasi Dokumen Untuk Identifikasi Konten E-Government*"Journal of Applied Intelligent System, Vol.1, No. 1, Februari 2016: 48-55

[3] . Alam Rahmatulloh, Rinaldi Munir.,"*Pencegahan Ancaman Reverse Engineering Source Code PHP Dengan Teknik Obfuscation Code Pada Extension PHP*" Conference Paper, Researchgate, October,2015.https://www.researchgate.net/publication/319454796

[4] . Fandy Ferdian Harryanto, Seng Hansun,"*Penerapan Algoritma C4.5 Untuk Memprediksi Penerimaan Calon Pegawai Baru Di PT WISE*",Jatisi, Vol. 3 No. 2 Maret 2017 ISSN: 2407-4322.

[5] . Febri Liantoni,"*Klasifikasi Daun Dengan Perbaikan Fitur Citra Menggunakan Metode K-Nearest Neighbor*",ULTIMATICS, Vol. VII, No. 2 ISSN 2085-4552 Desember 2015.

[6] . Slamet Pujiono, Armadyah Amborowati, M. Suyanto, "*Analisis Kepuasan Publik Menggunakan Weka DalamMewujudkan Good Governance Di Kota Yogyakarta*", Jurnal Dasi, Vol. 14 No. 2 Juni 2013, ISSN: 1411-3201.

[7] . Mohammad Saiful Islam Mamun, Mohammad Ahmad Rathore, Arash Habibi Lashkari, Natalia Stakhanova and Ali A. Ghorbani, "*Detecting Malicious URLs Using Lexical Analysis*", Network and System Security, Springer International Publishing, P467--482, 2016.

[8] . Michael Darling and Greg Heileman.,"*A Lexical Approach for Classifying Malicious URLs*"Computer Enginering and Cyber Security IEEE Publishing,978-1-4673-7813-0/15,2015

[9] . Ghulam Asrofi Buntoro.*,"Analisis Sentimen Calon Gubernur Jawa Timur 2018 Dengan Metode Naive Bayes Classifier"*JIPN (*Journal Of Informatics Pelita Nusantara*), Volume 4 No 1 Maret 2019 e-ISSN 2541-3724.

[10] . Syahfitri Kartika Lidya, Opim Salim Sitompul, Syahril Efendi.,"*Sentiment Analysis Pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (Svm) Dan K-Nearest Neighbor (K-Nn)*",Seminar Nasional Teknologi Informasi dan Komunikasi 2015 (SENTIKA 2015),ISSN: 2089-9815,Yogyakarta, 28 Maret 2015.

[11] . Tri Rochmadi.,"*Live Forensik Untuk Analisa Anti Forensik Pada Web Browser Studi Kasus Browzar*"Indonesian Journal of Business Intelligence,IJUBI - VOL. 1 NO. 1 (2018): 32 – 38.

[12] . UNB, "Canadian Institute for Cybersecurity" https://www.unb.ca/cic/datasets/url-2016.html

[13] . WEBSPAM-UK2007, dataset.http://chato.cl/webspam/datasets/uk2007

[14] . Wildan Budiawan Zulfikar, Nur Lukman,"*Perbandingan Naive Bayes Classifier Dengan Nearest Neighbor Untuk Identifikasi Penyakit Mata*"JOIN Volume 1 No. 2 Desember 2016,ISSN 2527-9165.